2N-3

人と計算機を情報資源とする統合情報検索システム CySearch の提案

三津石 智巳 †　望月 祥司 ‡　森嶋 厚行 ‡*

筑波大学 情報学群 知識情報・図書館学類 †　筑波大学大学院 図書館情報メディア研究科 ‡　JST さきがけ *

1. はじめに

各種情報資源から情報を入手するための手段として統合情報検索が注目されている．統合情報検索とは，複数の情報資源を対象とした情報検索を行い，その結果をひとつに統合して表示する機能である．例えば，Google[1)] や Bing[2)] 等は統合情報検索機能を持つ．

一方，Q&A サイトの例にも挙げられるように人は重要な情報資源である．しかし，これまで人と計算機情報資源に対する問合せのインターフェースは全く異なっていた．例えば，既存の Q&A サイトのシステムでは人を情報資源とする問合せと，計算機 (過去の質問・回答ログ) を情報資源とする問合せはそれぞれ別のインターフェースで行わなければならない．また，既存の統合情報検索は，検索対象が全て計算機上の情報資源であり，人を情報資源として含めた統合情報検索は行われていなかった．

そこで本研究では，人と計算機を情報資源とする統合情報検索システム CySearch (Cybernetic Search) の開発を行っている．人と計算機の統合情報検索の実現における第一の問題は，検索インターフェースの設計である．CySearch では，検索インターフェースとしてマイクロブログを用いる．その理由は，統合情報検索実現の一つのポイントに，各情報資源の応答時間の差を小さくすることがあるからである．例えば，ある人が，休日に T 市のレストランの検索を行ったとき，Web 情報資源からレストラン A が結果として返ってきたとする．その時，休日はレストラン A の定休日であるという情報がすぐに人から返ってくれば有益であるが，次の日に来ても役に立たない．論文[4)] によると，Q&A サイトにおいては 10 分未満で回答が得られる場合もある一方で，回答が得られるまで 2 時間以上かかった質問も全体の 3 割を占めている．したがって，一般の Q&A サイトのようなインターフェースは適切ではないと考えられる．一方，マイクロブログをインターフェースに用いた Q&A サービスである Q&A なう[3)] の回答平均時間は 36 秒 (2010/12/28 20:04 現在) である．このようにマイクロブログをインターフェースに用いることにより，人からの応答時間を早くすることができ，人と計算機の統合情報検索の応答時間の差を小さくすることができると予想できる．

第二の問題は統合情報検索の実現である．CySearch では，利用者のマイクロブログの発言をそのまま人 (回答候補者) に渡す一方で，Web や過去ログなどの計算機情報資源の検索のためには，発言からキーワード抽出をおこなって利用する．しかし，CySearch ではマイクロブログをインターフェースに採用し，かつ情報資源として計算機だけでなく人も含まれている．したがって，質問と回答 (もしくは検索と結果) のやりとりが，通常の Q&A サイトや情報検索と異なり，一般には Q と A のペアではなく会話のシーケンス (以下 QA シーケンス) となることが予想される．そこで，CySearch では，システムに記録されている発言列の中から QA シーケンスを抽出し，それを次のように計算機情報資源の検索に利用する．(1) Web 情報資源の検索においては，QA シーケンス中のキーワードを用いてクエリ拡張を行う．(2) 過去ログ検索においては，QA シーケンスを単位として保存，検索を行う．

2. 提案システム CySearch

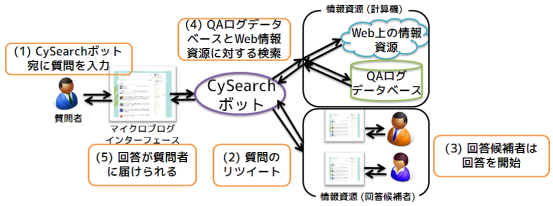

図 1　CySearch のシステム概要

図 1 は，CySearch のシステム概要である．CySearch の中心となるのは CySearch ボットである．CySearch ボットは，現在は Twitter ボットとして実装されている．CySearch ボットが対象とする計算機情報資源は，Web 検索エンジンと，CySearch ボットが受け付けたもしくは発言した全ての発言の過去ログデータベースである．また，CySearch ボットをフォローしている人々 (以下，回答候補者) が，CySearch における人の情報資源となる．質問者は，CySearch ボットをフォローしている必要はない．

CySearch は次のように動作する．質問者が，CySearch ボット宛に質問を投稿すると (図 1(1))，CySearch ボットはリツイートを行う (図 1(2))．CySearch ボットをフォローしている回答候補者は，そのメッセージに対して回答を開始する (図 1(3))．同時に，これまでのやりとり (QA シーケンス) が CySearch ボットの会話ログより抽出され，QA シーケンスを用いて QA ログデータベースと Web 情報資源に対しても検索が行われ (図 1(4))，その結果と共に，回答は，CySearch ボットから質問者に届けられる (図 1(5))．一般には，質問者は必要な回答が得られるまで会話を続ける．

3. CySearch の設計

本節では，まず，QA シーケンスについて説明し，次に統合情報検索の手順について説明する．

3.1 QA シーケンス

これまで述べたように，質問者・回答者から CySearch ボットに入力された全てのメッセージ (発言) をシステムは過去

CySearch: Proposal of a System for the Blended Search of Human and Computer Information Sources
Tomomi Mitsuishi† Shoji Mochizuki‡ Atsuyuki Morishima‡*
College of Knowledge and Library Sciences, School of Informatics Univ of Tsukuba.†
Grad. Sch. of Library, Information and Media Studies, Univ. of Tsukuba.‡
PRESTO, JST*

```
01. // 条件分岐：M_i が質問開始メッセージであるかそれ以外であるか
02. if (M_i.isQuestion) { // 質問開始メッセージ
03.     autoReplyToQuestioner(M_i)
04.     retweet(M_i)
05. } else { // 質問開始メッセージ以外
06.     // 条件分岐：M_i の発言者が回答候補者か質問者か
07.     if (M_i.isFromRespondents) { // M_i の発言者は回答候補者
08.         deliverToQuestioner(M_i)
09.     } else {
10.         autoReplyToQuestioner(M_i)
11.         deliverToTheRespondent(M_i)
12.         retweet(M_i)
13.     }
14. }
15. MessageList.add(M_i)
```

図 2 CySearch の処理概要

メッセージログとして保存しており，そこから，QA シーケンスを発見し，検索に利用する．

まず，CySearch ボットが管理しているメッセージログを説明する．これは，質問者・回答者からボットに行われた全ての発言 M_i の列を保存した物であり，時系列に並べたリスト $MessageList = [M_1, M_2, \cdots, M_n]$ として表現できる．ただし，各 Mi は次の 5 つ組である．

$$M_i = (id_i, user_i, reply\_to_i, reply\_to\_user_i, text_i)$$

ここで，id_i はメッセージ M_i の id である．$user_i$ は発言者の id である．$reply\_to_i$ は M_i の返信先のメッセージの id である．$reply\_to\_user_i$ はメッセージの宛先の人の id である (特定の宛先が無い場合には NULL とする)．$text_i$ は M_i のメッセージの内容を表す文字列である．

この $MessageList$ と，あるメッセージ $M_i (1 \leq i \leq n)$ が与えられたとき，M_i に関する QA シーケンス qa_i を次のように定義する．

$$qa_i = [M_{i_0}, M_{i_1}, \cdots, M_{i_k}]$$

ただし，$i_k = i$ であり，qa_i は次の 3 つの条件を満たすものである．(1) $reply\_to_{i_0} = null$, (2) $reply\_to_{i_j} = id_{i_{j-1}}$, (3) $reply\_to\_user_{i_{j-1}} = u$ ならば $user_{i_j} = u$．ここで，条件 (1) は，質問者からの質問は他の発言へのリプライでないということである．このメッセージを「質問開始メッセージ」と呼ぶ．条件 (3) は，質問者に対するボットからの発言に対しては，質問者が答えていなければならないという制約を表す．

3.2 統合情報検索

まず，統合情報検索の全体像について説明し，次に Web 情報資源の検索と回答ログ検索について説明する．

全体像．CySearch の処理概要を表す擬似コードを図 2 に示す．入力は，最新の発言 M_i である．

疑似コードの処理を次に説明する．まず，M_i が質問開始メッセージであるか，それ以外であるかを判定する (2 行目)．M_i が質問開始メッセージであれば，次の処理を行う．まず，過去ログデータベースを検索しその結果を質問者に返す (3 行目)．次に，M_i のリツイートを行う (4 行目)．一方，M_i が質問開始メッセージでない場合 (回答など) には，M_i の発言者が回答候補者か，それとも質問者なのかを判定する (7 行目)．回答候補者の発言であれば，統合情報検索結果を構築し，質問者に通知する (8 行目)．回答者の発言にリプライする質問者の発言であれば，まずは過去ログデータベースの検索して結果を返し (10 行目)，さらに，リプライされた回答者への通知，宛先を決めないリツイート (回答候補者全員が見ることができる) の順に処理を行う (11-12 行目)．最後に，いずれの場合も M_i を保存する (15 行目)．

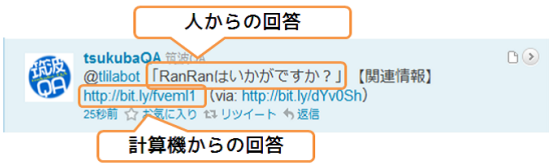

図 3 統合情報検索結果の提示

統合情報検索の結果は，人からの回答と，計算機からの応答を示す関連情報 URL をあわせて表示する (図 3)．

Web 情報資源の検索． 図 2 の疑似コードの 8 行目 deliverToQuestioner(M_i) の中で，CySearch ボットは Web 情報資源の検索を行う．その際，M_i だけでなく，QA シーケンスを用いてクエリ拡張を行う．具体的な手順を次に示す．

(1) M_i に関する QA シーケンス qa_i を抽出する．
(2) qa_i の全てのメッセージからキーワード抽出を行う．
(3) 抽出されたキーワードを結合して検索クエリを作成し，Web 情報資源の検索を行う．

回答ログ検索． 図 2 の疑似コードの 3 行目と 10 行目の autoReplyToQuestioner(M_i) の中で，CySearch ボットは QA シーケンスを利用し，次の手順で過去ログ検索を行う．

(1) M_i に関する QA シーケンス qa_i を抽出する．
(2) $MessageList$ より，次の条件を満たす $qa_{l \neq i}$(l は複数存在) を抽出する．すなわち，qa_i の全ての構成要素 M_{i_j} に関して $sim(text_{i_j}, text_{l_j}) > \theta$．ここで，$sim(t, t')$ は t と t' の類似度，θ は閾値である．
(3) 抽出された各 qa_l の最後のメッセージ M_l に対するリプライの集合 $ReplyMsg(M_l)$ を，質問者に通知する．複数 qa_l が存在する場合には全てを通知する．

4. まとめと今後の課題

本稿では，人と計算機を情報資源とする統合情報検索システム CySearch の提案を行った．本システムでは，質問者と，Web 情報資源や回答候補者，過去ログデータベースとのやりとりを CySearch ボットが仲介するが，その仲介の設計によってシステムの振る舞いが大きく変わることが予想される．今後の課題は，そのような設計に関するより詳細な検討，および実験による評価等が挙げられる．

5. 謝　　辞

本研究の一部は科学研究費補助金若手研究 (B)(#20700076) および JST さきがけ「情報環境と人」の支援による．

参　考　文　献

1) "Google". http://www.google.co.jp/.
2) "Bing". http://www.bing.com/.
3) "Q&A なう". http://qa-now.com/.
4) 辻慶太, 楳原衣恵, 木川田朱美. "Q&A サイトと公共図書館レファレンスサービスの質問回答力比較：正答率を中心として". 日本図書館情報学会春季研究集会発表要綱, 日本図書館情報学会, 2009, p. 79-82. http://hdl.handle.net/2241/102741, (参照 2010-12-28).